2.4 Crosstabulations and Scatter Diagrams

* Crosstabulations and scatter diagrams are used to summarize data in a way that reveals the relationship between two variables.

* Crosstabulation is a tabular summary of data for 2 variables.

Example: Consider the following quality rating and meal price for 300 restaurants:

Restaurant	Quality Rating	Meal price ($)
1	Good	18
2	Very Good	22
3	Good	28
4	Excellent	38
5	Very Good	33
⋮	⋮	⋮
300	Good	13

Each resturante provides a quality rating and a meal price.

a) Construct a crosstabulation for the data

b) Develop a relative and percent frequency distribution for quality rating

c) " " " " " " " = meal price

d) construct a row percenteges for each quality rating category

d) What is the relationship between the quality rating and meal price?

e) " " " column " " " meal price

* Quality rating: is qualitative variable with categories: good, very good, excellent.

* Meal price: is quantitative variable that ranges from 10 $ to 49 $

Quality rating	Meal Price $10-19	$20-29	$30-39	$40-49	Total
Good	42	40	2	0	84
Very Good	34	64	46	6	150
Excellent	2	14	28	22	66
Total	78	118	76	28	300

Column ⇓ ; → row total ; row ⇒ ; column total ; 300 → crosstabulation

* For example restaurant 5 provides a quality rating very good with meal price $33. This restaurant belongs to the cell in row 2 and column 3.

- The greatest number of restaurants in the sample is 64 have a very good rating and meal price in $20 – 29 range.
- Only 2 restaurants with excellent rating and meal price in $10 – 19 range.

(b)

Quality rating	Relative frequency	Percent frequency
Good	$\frac{84}{300} = 0.28$	$0.28 \times 100 = 28$
Very Good	$\frac{150}{300} = 0.50$	$0.50 \times 100 = 50$
Excellent	$\frac{66}{300} = 0.22$	$0.22 \times 100 = 22$
	Total = 1.00	Total = 100

28% of the restaurant were rating good.
50% " " " = very good.
22% " " " = excellent.

(c)

Meal Price	Relative frequency	Percent frequency
$10 – 19	$\frac{78}{300} = 0.26$	$0.26 \times 100 = 26$
$20 – 29	$\frac{118}{300} = 0.39$	$0.39 \times 100 = 39$
$30 – 39	$\frac{76}{300} = 0.25$	$0.25 \times 100 = 25$
$40 – 49	$\frac{28}{300} = 0.09$	$0.09 \times 100 = 9$
	Total = 1.00	Total = 100

26% of the meal price are in the lowest class $10 – 19
39% " " " " " " " = next higher class and so on

(d) Higher meal prices are associated with the higher quality rest.
(f) lower meal prices = = = lower = =

→ to know the relationship between the two variables within cross tabulation, we convert the entries into row percentages or column percentages.

(d)

Quality Rating	Meal price $10-19	$20-29	$30-39	$40-49	Total
Good	$\frac{42}{84}\times100=50$	$\frac{40}{84}\times100=47.6$	$\frac{2}{84}\times100=2.4$	$\frac{0}{84}\times100=0.0$	100
Very Good	$\frac{34}{150}\times100=22.7$	$\frac{64}{150}\times100=42.7$	$\frac{46}{150}\times100=30.6$	$\frac{6}{150}\times100=4.0$	100
Excellent	$\frac{2}{66}\times100=3.0$	$\frac{14}{66}\times100=21.2$	$\frac{28}{66}\times100=42.4$	$\frac{22}{66}\times100=33.4$	100

↑ Row percentages for each quality rating category

* For the lowest quality restaurant (good), we see the greatest percentages are for the less expensive restaurants (50% have $10-19 meal prices and 47.6% have $20-29 meal prices)...
* For the greatest quality restaurants (excellent), we see the greatest percentages are for the more expensive restaurants (42.4% have $30-39 meal prices and 33.4% have $40-49 meal prices).

(e)

Quality Rating	$10-19	$20-29	$30-39	$40-49
Good	$\frac{42}{78}\times100=54$	$\frac{40}{118}\times100=34$	$\frac{2}{76}\times100=3$	$\frac{0}{28}\times100=0$
Very Good	$\frac{34}{78}\times100=44$	$\frac{64}{118}\times100=54$	$\frac{46}{76}\times100=61$	$\frac{6}{28}\times100=21$
Excellent	$\frac{2}{78}\times100=2$	$\frac{14}{118}\times100=12$	$\frac{28}{76}\times100=36$	$\frac{22}{28}\times100=79$
	100	100	100	100

↑ Column percentages for each ↓ category meal price

Simpsons Paradox

Conclusions drawn from two or more separate cross tabulations that can be reversed when data are aggregated into a single cross tabulation.

Example: Consider the following two cross tabulations

Cross tabulation for School 1

Gender	10th class	5th class	
M	29 (91%)	100 (85%)	129
F	3 (9%)	18 (15%)	21
	32 (100%)	118 (100%)	150

Cross tabulation for school 2

Gender	10th class	5th class	
M	90 (90%)	20 (80%)	110
F	10 (10%)	5 (20%)	15
	100 (100%)	25 (100%)	125

Simpson's Paradox:

Gender	School 1	School 2	
M	129 (86%)	110 (88%)	239
F	21 (14%)	15 (12%)	36
	150 (100%)	125 (100%)	275

In Simpson's paradox, we need to be careful, when drawing conclusions using aggregated data.

Hidden variable is 10th class and 5th class.

Scatter Diagram and Trendline

- A scatter diagram is a graphical presentation of the relationship between two quantitative variables.
- Trendline: is a line that provides an approximation of the relationship.

Example: The following 10 observation are for two quantitative variables
x: number of commercials
y: sales ($100s)

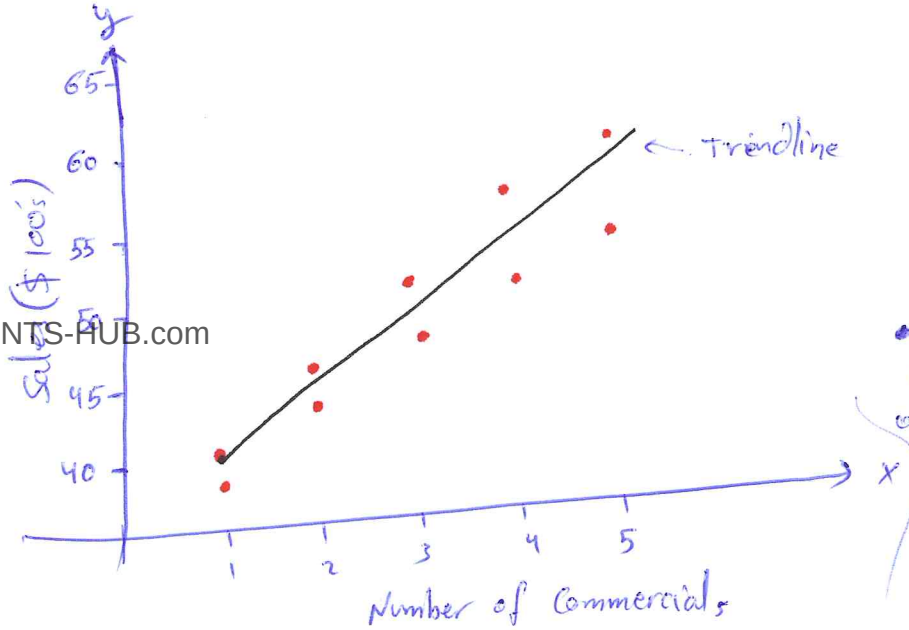
Number of Commercials (x)	Sales ($100s) y
2	50
5	57
1	41
3	54
4	54
1	38
5	63
3	48
4	59
2	46

a) Develop a scatter diagram for the relationship between x and y.

b) What is the relationship, if any, between x and y?

(c) Is the relation perfect?

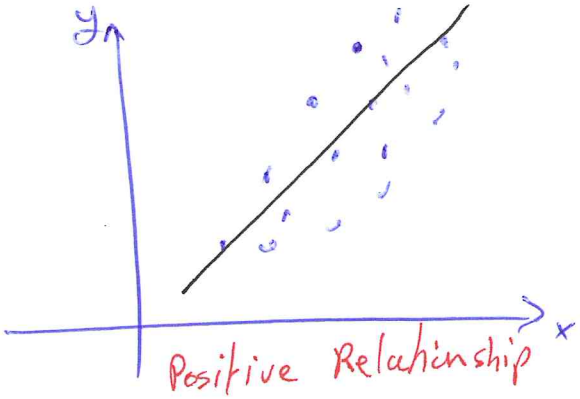
(a)



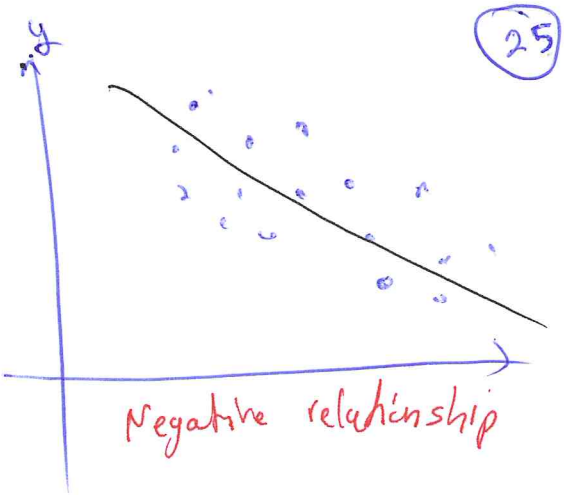
(b) The scatter diagram indicates a positive relationship between x and y.

- Higher sales associated with higher number of commercials.

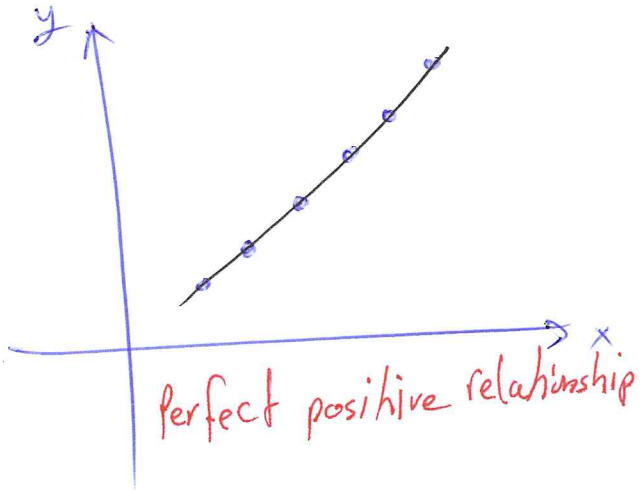
(c) The relation is not perfect. Because not all the points are on the trendline



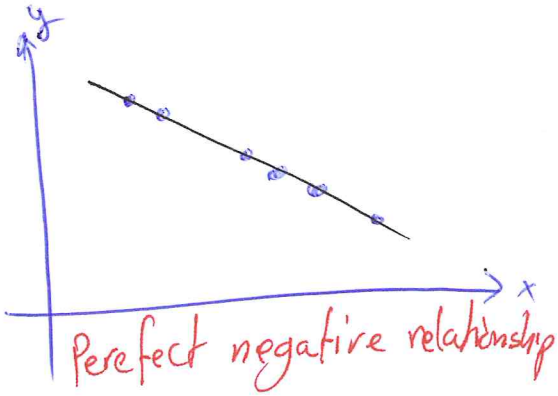
Positive Relationship



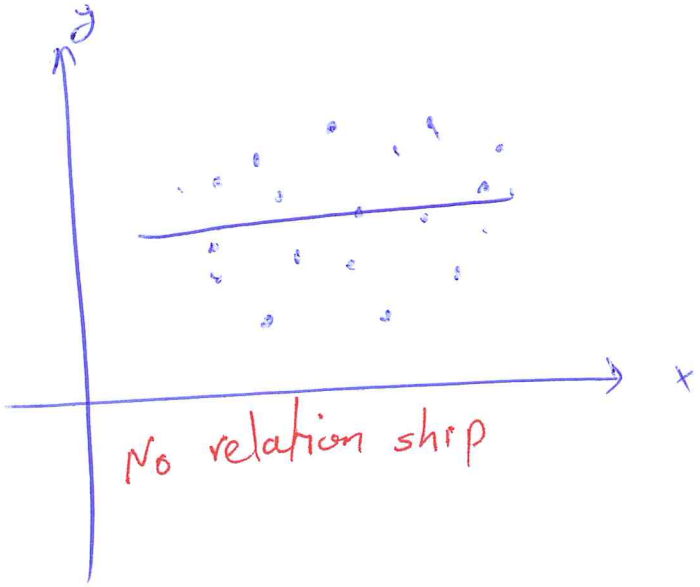
Negative relationship



Perfect positive relationship



Perefect negative relationship



No relation ship